Data 583 Life Expectancy (WHO)

Justin Chan, Kenny Tong, Viji Rajagopalan

22 Mar, 2023

Data 583 Life Expectancy - Final Report (Life Expectancy Data)

1. Introduction and Hypotheses

Life expectancy has always been an area of interest for humanity. The goal of this project is to study a dataset that contains information on life expectancy and identify some of the variables that significantly impact life expectancy. The dataset chosen for the study has life expectancy data of 193 countries between 2000-2015, together with different predictive factors. Broadly speaking, predicting variables are categorized into 4 major areas : Immunization, Mortality, Economical, and Social, containing a total of 21 individual variables. Our hypothesis is that a subset of variables from this dataset would be able to explain and predict life expectancy with good accuracy (say > 80%). The dataset has a mix of variable types – continuous and discrete. Within discrete types, some variables are ordinal, and some are non-ordinal or nominal. With such a mix and complexity of data, we also hypothesize that all variables will not share a simple linear relationship with the predictor variable and modelling of life expectancy will require a more complex model. We analyze and validate several statistical models throughout the report with the primary goal of identifying an adequate model for the dataset.

2. Dataset overview

Variables Summary and Categories

Life expectancy is the response variable in this dataset. This represents the mean life expectancy (in age) by specific country and year combination. Refer Figure-1 below for the list of predictor variables and their categories. The dataset also had 2563 missing values in various columns and ..... the NA values are generally imputed by the respective column mean. Percentage expenditure variable is removed from the entire assessment as the values present in this column are unclear. Another variable country is removed for the purpose of studying life expectancy globally. The resulting dataset with x rows and y columns are then studied closely to understand their correlation effects with the response variable life expectancy. #(Kenny, please add if any other deletion done in .... line and fill in on x and y)

Variable	Unit of Measurement/Data Category	Continuous vs Discrete	Variable	Unit of Measurement/Data Category	Continuous vs Discrete
Life Expectancy	Years Old (Age)	Continuous	Total expenditure	Percentage	Continuous
Country	Nominal Data	Discrete	Percentage expenditure	Percentage	Continuous
Year	Ordinal Data	Discrete	GDP	Currency (USD)	Continuous
Status	Nominal Data	Discrete	Population	Count	Discrete
Adult Mortality	Count Data	Discrete	Income composition of resources	Percentage	Continuous

Variable	Unit of Measurement/Data Category	Continuous vs Discrete	Variable	Unit of Measurement/Data Category	Continuous vs Discrete
Infant deaths	Count Data	Discrete	Schooling	Mean (Years)	Continuous
Under-five deaths	Count Data	Discrete	Alcohol	Litres	Continuous
Hepatitis B	Percentage	Continuous	HIV/AIDS	Percentage	Continuous
Measles	Count Data	Discrete	BMI	Average BMI	Continuous
Polio	Percentage	Continuous	Thinness 1-19 years	Percentage	Continuous
Diphtheria	Percentage	Continuous	Thinness 5-9 years	Percentage	Continuous

Initial analysis using linear regression

Life expectancy is a continuous variable and the first choice is building a linear regression model which is simple and interpretable. A BIC backward step model variable selection method is also applied on the full model to arrive at a parsimonious model containing only significant predictor variables. Following table Table A provides a summary of the two models.

Models	No. of Variables	AIC Score	Adj R-squared Score
Original Model	20	7642.14	0.8299
Reduced Model	12	7604.24	0.8296

The number of independent variables are now effectively reduced to 12, together with a lower AIC score of 7604.34. Meanwhile, the adjusted R-squared score is well kept at nearly the same level as in the original model. The reduced

model is able to explain more than 82% of variation in the response variable and its performance is above the anticipated 80%.

The reduced model now contains the following 12 variables : Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + GDP + thinness..1.19.years + Income.composition.of.resources + Schooling.

With performance of the model over 80%, the next step is to look at the error diagnostics from the model.

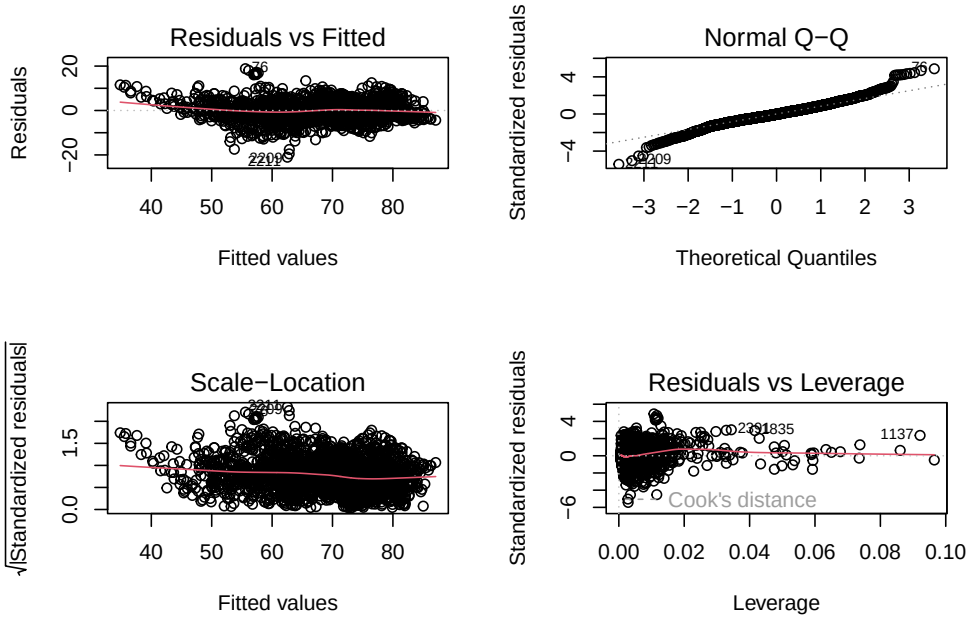
We eliminate the Status variable from the selected list of variables in the reduced model as this is a factor variable with two statuses and not continuous. We first study the effect of the model without this variable.

3. Regression Analysis

Linear model and diagnostics

The initial model shows that we are able to explain approximately 82% of variability of our response variable using the selected predictor variables. The next step is to look at the error diagnostics from the model.

```
par(mfrow=c(2,2))
plot(lmmod2)
```



The QQ plot suggests that the model is heavy tailed and the data on both ends of the quantiles do not fit on a straight line. This is an indication that the current linear regression based model is not fitting the data well. Based on this, we undertake some additional testing to validate if the model is adequate and valid.

a. Life expectancy variable distribution From the histogram, it can be noticed that the response variable life expectancy is not normally distributed. From the plot, it also seems like a bimodal distribution of life expectancy data in the dataset.

b. Normal distribution test for our y variable Next, we evaluate to confirm if the response variable is normally distributed using Shapiro-Wilk test. The test has a p-value that is very small and is less than 0.05, this indicates that our response variable if not normally distributed.

```
#Shapiro-Wilk Test

shapiro.test(df$Life.expectancy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Life.expectancy
## W = 0.95676, p-value < 2.2e-16
```

```
#Finding: Since df$Life.expectancy p-value is less than .05, indicate that our y variable is not normally
```

As response is not normal, the next step is to validate with a hypothesis test for validating correct specification of parametric MLR models.

c. Parametric model specification test Another test to see if the above parametric model specification is correct.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.2.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(lmmod2)
```

```
##
##  RESET test
##
## data:  lmmod2
## RESET = 121.2, df1 = 2, df2 = 2763, p-value < 2.2e-16
```

d. Consistent nonparametric inference

```
##
## Consistent Model Specification Test
## Parametric null model: lm(formula = Life.expectancy ~ Adult.Mortality +
##                           infant.deaths + Hepatitis.B + BMI + under.five.deaths
##                           + Polio + Diphtheria + HIV.AIDS + GDP +
##                           thinness..1.19.years +
##                           Income.composition.of.resources + Schooling, data =
##                           df, x = TRUE, y = TRUE)
```

```
## Number of regressors: 12
## IID Bootstrap (399 replications)
##
## Test Statistic 'Jn': 21.17521    P Value: < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Null of correct specification is rejected at the 0.1% level
```

All the diagnostic tests indicate that linear regression is not an appropriate model for the given data as assumptions for the model are violated.

Parametric regression models and relative assessments

As the linear model is not adequate, we move on to model this with other models that do not assume normal distribution. The models selected for the given dataset are LASSO and Neural Net with linear activation function. The following variables are selected for rest of the modeling based on correlation of the variables with the response variable and our knowledge on the domain. Here is a summary of the variable selection and our comments.

Data Categories	**Vaiables**
Economical Data	Total expenditure, Percentage expenditure, GDP, Income composition of resources
Social Data	Country, Status, Population, Schooling, Alcohol, BMI, Thinness 1-19 years, Thinness 5-9 years
Mortality Data	Adult Mortality, Infant deaths, Under-five deaths
Immunization Data	Hepatitis B, Measles, Polio, HIV/AIDS, BMI, Diphtheria

Column Name	**Type**	**LM**	**LASSO**	**NN**	**NPREG**	**Reason of Removal**
Country	(Discrete)					Since we wanted to build models for all countries
Year	(Discrete)					ordinal type data and based on domain knowledge, not consider important
Status	(Discrete)					nominal type data and based on domain knowledge, not consider important
Adult Mortality	(Discrete)	X	X	X	X	
Infant deaths	(Discrete)	X	X	X	X	
Under-five deaths	(Discrete)	X	X	X	X	
Hepatitis B	(Continuous)	X	X	X	X	
Measles	(Discrete)					Since it is a count and discrete type data and weak correlation with our predictor
Polio	(Continuous)	X	X	X	X	
Diphtheria	(Continuous)	X	X	X	X	
Total Expenditure	(Continuous)					based on domain knowledge, not consider important
Percentage Expenditure	(Continuous)					based on domain knowledge, not consider important
GDP	(Continuous)	X	X	X	X	
Population	(Discrete)					no correlation with our predictor indicated by our correlation plot
Income composition of resources	(Continuous)	X	X	X	X	

Column Name	Type	LM	LASSO	NN	NPREG	Reason of Removal
Schooling	(Continuous)	X	X	X	X	
Alchol	(Continuous)					based on domain knowledge, not consider important
HIV/AIDS	(Continuous)	X	X	X	X	
BMI	(Continuous)	X	X	X	X	
Thinness 1-19 years	(Continuous)	X	X	X	X	
Thinness 5-9 years	(Continuous)					range already covered in 1-19 Thinness 1-19 years
status.val	(Continuous)					based on domain knowledge, not consider important

Two different supervised algorithms tried on the dataset. They do not have the constraint of a normal distribution for response variable.

First did a train and test split so we can measure the MSE and compare how each of the models are performing in terms of minimizing MSE.

PRESS comparison for the three models

	LM	LASSO	NN
PRESS	15.93367	15.98972	22.43056

Test R2 comparison for the three models

	LM	LASSO
R2	0.829139273435324	0.829061931452822

As we compare linear model, lasso and neural net, we see that the test MSE is minimum for LASSO model. So this is a model that can be considered for the dataset.

Diagnostics

Nonparametric regression

The response variable shows a bimodal distribution and nonparametric regression performs better on such datasets per literature. We next try non parametric regression on the dataset.

```
library(np)
# n <- names(df)
# f <- as.formula(paste("df$Life.expectancy ~", paste(n[!n %in% "Life.expectancy"], collapse = " + ")))
#
# model_np <- npregbw(Life.expectancy ~ Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.d

# model_np <- npreg(bws = model_np)
# summary(model_np)
model_np <- readRDS("model_np.rds") #PreTrained Model
summary(model_np)
```

Diagnostics

```
##
## Regression Data: 2778 training points, in 12 variable(s)
##                Adult.Mortality infant.deaths Hepatitis.B      BMI
## Bandwidth(s):       389457535       6733757   225092161 79216285
##                under.five.deaths   Polio Diphtheria HIV.AIDS          GDP
## Bandwidth(s):          95308072 5825954   19248839 1.393258 167351562078
##                thinness..1.19.years Income.composition.of.resources Schooling
## Bandwidth(s):             37667667                         1071202  15165344
##
## Kernel Regression Estimator: Local-Linear
## Bandwidth Type: Fixed
## Residual standard error: 3.345092
## R-squared: 0.8722143
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 12
```

We see that the R^2 is increased to 87% approximately. Done with local linear estimator and cv.aic. This is a cross validated model and help estimate the long run performance. Can we see BIC?

```
#npsigtest_npreg <- npsigtest(model_np)    #10 HRs to run...
```

```
> npsigtest(model_np)

Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications, Pivot = TRUE, joint = FALSE)
Explanatory variables tested for significance:
Adult.Mortality (1), infant.deaths (2), Hepatitis.B (3), BMI (4), under.five.deaths (5), Polio (6), Diphtheria (7), HIV.AIDS (8),
 GDP (9), thinness..1.19.years (10), Income.composition.of.resources (11), Schooling (12)

             Adult.Mortality infant.deaths
Bandwidth(s):       389457535       6733757
             Hepatitis.B      BMI under.five.deaths
Bandwidth(s):   225092161 79216285          95308072
               Polio Diphtheria HIV.AIDS
Bandwidth(s): 5825954   19248839 1.393258
                    GDP thinness..1.19.years
Bandwidth(s): 167351562078             37667667
             Income.composition.of.resources
Bandwidth(s):                         1071202
             Schooling
Bandwidth(s):  15165344

Individual Significance Tests
P Value:
Adult.Mortality                 < 2e-16 ***
infant.deaths                   < 2e-16 ***
Hepatitis.B                     0.047619 *
BMI                             < 2e-16 ***
under.five.deaths               < 2e-16 ***
Polio                           < 2e-16 ***
Diphtheria                      < 2e-16 ***
HIV.AIDS                        < 2e-16 ***
GDP                             < 2e-16 ***
thinness..1.19.years            < 2e-16 ***
Income.composition.of.resources < 2e-16 ***
Schooling                       < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: npsigtest_npreg result

We measure the significance of the variables for a parsimonious model. All the parameters used are significant. Summarizing the different models and the performance assessed during the course of this project # Add table

Model improvements

1. status variable and other ordinal, nominal variables

2. train and test split for np, but what we have now is sufficient for measuring long run performance
3. multicollinearity between the variables infant.deaths and under.five.deaths. Remove one of the variables and study if there are improvements in performance
4. measure without imputing data

Challenges

1. 30 hours for npreg
2. 30+ hours for model significance
3. null values in the dataset, columns dropped - , columns imputed with mean -

Conclusion

np and LASSO are suitable for this dataset. We find that life expectancy is....

Appendix

Checking for Multicollinearity

As Multicollinearity can potentially affect the accuracy of regression model and we have 22 variables, a correlation study is undertaken to understand and assess the situation. A correlation plot has identified a number of correlation problems. It is found that infant deaths and under.five.deaths are nearly 100% correlated. The relation between the deaths rates of the two close age groups is easily interpretable. In addition, there are three heavily correlated pairs which is defined by the abs(correlation coefficient)>0.7 between the variables. They include (a) (immunization rate of) 'Polio'-vs-'Diphtheria', (b) 'income composition of resources'-vs-'Schooling', and (c) between the two thinness measures for the age groups 5-9 vs 10-19. Pairs (a) and (c) are justifiable while the relation for (b) demonstrate a relatively subtle relation. Other than that, the degree of multicollinearity is acceptable and not too worrying.